Mapping the News: COVID-19 in the NY/NJ Metro Area

by Tammy Ann Syrek-Marshall

I. Introduction

a. Background

Epidemics and pandemics have occurred throughout human history. From the plagues of ancient Greece and Rome to the more recent Ebola Virus, these outbreaks have been responsible for the deaths of thousands if not millions.

In the 21st century, along with advances in science and medicine, we still look to the leadership of our governments to take responsibility for mitigating the effects of these epidemics as well as for keeping the public informed. This information often takes the form of local and national news reports.

On December 31, 2019 the first reports of a new strain of pneumonia afflicting dozens of people in Wuhan China made the headlines in the New York Times as well as other news outlets.

On January 7, 2020 this virus was identified as a novel, previously unknown strain.

The first death from this virus was reported on January 11, 2020.

By January 21, 2020 the first case of this Coronavirus, officially named COVID-19, was reported In Washington State.

As of March 1, 2020, while the virus continued to spread quickly to other countries, New York and New Jersey began seeing their first cases of COVID-19.

b. Problem and Discussion

In this report, the primary aim is to provide government officials, community leaders, as well as business and news media executives with a methodology for understanding through data visualization information dissemination during a crisis situation.

While the outbreaks are the major focus of the news, the effects of this virus can also be seen in other community related stories.

Each state in the U.S. is an entity onto its' self, with differences in their governments and their communities. With that in mind, it might be possible to gain new insights into the effects of this epidemic on society by doing a comparative analysis of the news coverage of New Jersey to the coverage of New York.

The more information on the effects of a society-wide disruptor, the more efficient and effective the planning and problem solving can be.

II. Data

a. Overview

By using the tools of data science and geospatial data, I propose to map out community specific news articles based on the geographical focus of their stories.

In order to create the Maps, I will use Folium library to visualize the data.

As for the data itself, the News Article data will be derived from searching databases such as ProQuest then compiling the data into an excel spreadsheet, formatted so as to be transformable into a Pandas dataframe.

Additional community data might be uploaded from FourSquare, if more information is needed. It is also possible to obtain datasets from sources such as the CDC and Census. However, I expect this might not be required.

Once the data is visualized, additional information will be provided in the final product relating to each marker on the map.

As an example, a story may cover the cancellation of a public event, much like the South by Southwest (SXSW) event in Austin Texas, due to concerns over the Coronavirus. The FourSquare data can provide additional information about the venue as well as nearby venues also affected by the cancellation.

III. Methodology

a. Data Collection

Data collection for this project consisted primarily in search through Newspaper Archives and Databases for relevant articles. Once articles were identified, the citations were aggerated together using a citation manager, in this case Zotero.

Specific limitations were placed on article searches and selections. These limitations focused on relevance to the project, location, timeframe, and a clear reference to a geographical location.

b. Data Processing

Once a sufficient number of relevant articles had been identified and their citations uploaded into Zotero, the next step was to clean the data. Over the course of several days the citations were examined for missing and inaccurate data. Citations that did not meet the criteria were removed. Then geographical data was added in preparation for the inclusion of geographical coordinates.

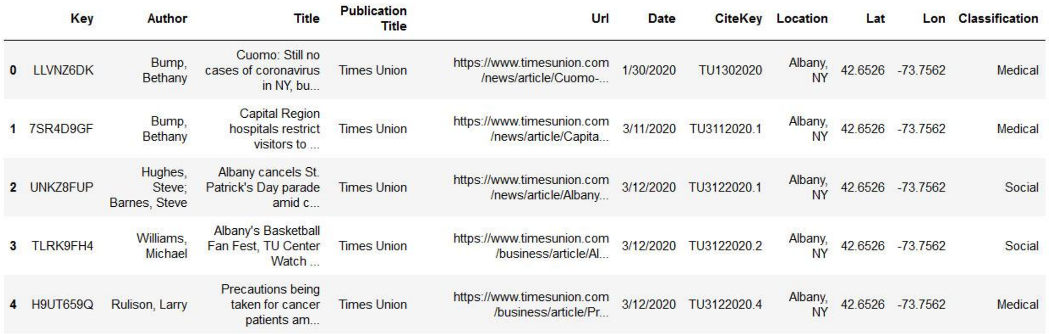
c. Conversion to Datafile

Once the citation data was cleaned in Zotero, it was downloaded into an Excel csv file. At this time additional processing was done. In addition to reviewing the data, CiteKeys and Coordinates were added to each field. Articles that covered more than one geographical location were split and assigned related CiteKeys.

Additionally, the articles were divided up into two groupings. The Medical Dataset consisted of articles that focused on health and medical issues related to the COVID-19 virus. Issues such as rates of infection, new cases identified, hospitalizations, and testing. The Social Dataset comprised articles that focused on the impact to society of the Pandemic. These include cancelation of events, closing of schools and other venues, as well as the effects this has on people and society.

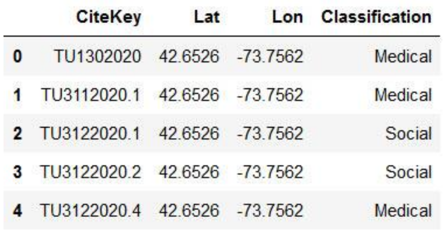
d. Mapping the Data

With the Notebook already set up, the first step was importing the needed, or potentially needed libraries. The next step involved uploading the three csv files and converting them to Pandas Dataframes.



	Key	Author	Title	Publication Title	Url	Date	CiteKey	Location	Lat	Lon	Classification
0	LLVNZ6DK	Bump, Bethany	Cuomo: Still no cases of coronavirus in NY, bu...	Times Union	https://www.timesunion.com /news/article/Cuomo-...	1/30/2020	TU1302020	Albany, NY	42.6526	-73.7562	Medical
1	7SR4D9GF	Bump, Bethany	Capital Region hospitals restrict visitors to ...	Times Union	https://www.timesunion.com /news/article/Capita...	3/11/2020	TU3112020.1	Albany, NY	42.6526	-73.7562	Medical
2	UNKZ8FUP	Hughes, Steve; Barnes, Steve	Albany cancels St. Patrick's Day parade amid c...	Times Union	https://www.timesunion.com /news/article/Albany...	3/12/2020	TU3122020.1	Albany, NY	42.6526	-73.7562	Social
3	TLRK9FH4	Williams, Michael	Albany's Basketball Fan Fest, TU Center Watch ...	Times Union	https://www.timesunion.com /business/article/Al...	3/12/2020	TU3122020.2	Albany, NY	42.6526	-73.7562	Social
4	H9UT659Q	Rulison, Larry	Precautions being taken for cancer patients am...	Times Union	https://www.timesunion.com /business/article/Pr...	3/12/2020	TU3122020.4	Albany, NY	42.6526	-73.7562	Medical

For each dataframe, slicing was used to create a new dataframe comprised of only four columns. These columns, CiteKey, Lat, Lon, and Classification, are the ones needed to create the maps.



	CiteKey	Lat	Lon	Classification
0	TU1302020	42.6526	-73.7562	Medical
1	TU3112020.1	42.6526	-73.7562	Medical
2	TU3122020.1	42.6526	-73.7562	Social
3	TU3122020.2	42.6526	-73.7562	Social
4	TU3122020.4	42.6526	-73.7562	Medical

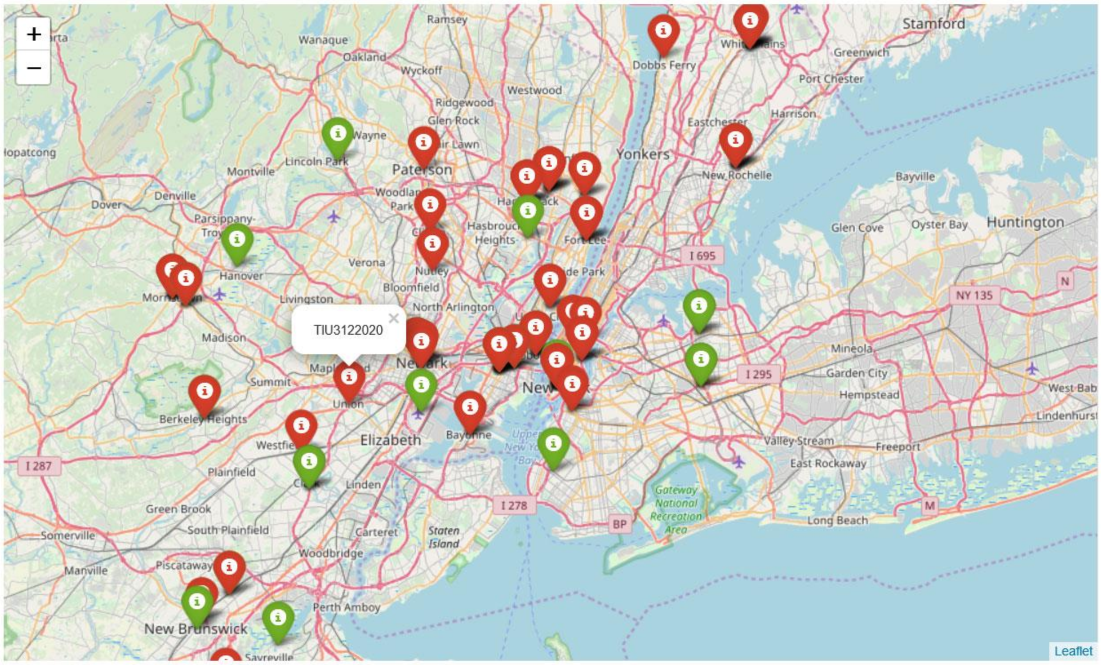
From these smaller Dataframes, a Location List was created using the Lat and Lon Coordinates.

It was then that the maps could be created using Folium. The first map was derived from the entire Dataset, with each point marked in blue. The second map comprised just those data fields with the Classification of Medical, with markers in red. Then a map was created using the Classification subset of Social, using green markers. The final map combined the two subsets into a single map, with both the red markers and the green markers.

Each marker is labeled with the CiteKey. The CiteKey is tied to the original citations for each article. These citations will be made available in the presentation and once the project is posted online.

IV. Results

The result was the creation of an interactive map that plotted the effects of the COVID-19 Pandemic on the New York/New Jersey area as reported by regional newspapers.



The Citation Key refers to a specific article. In the case of the Marker identified on the map above, the following would be the related citation.

(TIU3122020) Cryan, Kathy. 2020. "Union Businessman Diagnosed with Coronavirus." *TapIntoUnion*, March 12, 2020. https://www.tapinto.net/towns/union/articles/union-businessman-diagnosed-with-coronavirus

Once I regain full access to my Jupyter Notebook on Watson Studio, I plan on saving the map files. Then I hope to be able to post an interactive report on my blog, as well as a related report on LinkedIn.

V. Discussion

It is important to mention that this is only a sampling of newspaper articles from the New York/New Jersey area. It is intended as a proof of concept. An example of how mapping News Reports can illustrate patterns of change in society over time.

The time frame runs from Late January 2020 to March 15, 2020, yet the Pandemic is still ongoing. Given the time and resources, it would have been more illustrative to also map patterns based on the dates of the articles.

Another way to expand on this would have been to extend coverage to include broadcast news reports as well as print and online news.

I did explore the possibility of linking the articles directly to the map markers, but found that to be beyond my current skill level.

One thing I chose not to do was to include FourSquare data. After consideration, I decide it would only make the make the maps cluttered and confusing. Conceptually, however, it would have been useful to map the venues in the areas surrounding each geographical location. To show which businesses might have been negatively affected by the outbreak. Perhaps that, in of itself is worth a separate project.

VI. Conclusion

Data analysis and Mapping are important tools to use in understanding the progression of a Pandemic such as COVID-19. Most that have been done already focus on the disease. Mapping out the affects that a pandemic has on society is also a valuable tool, one that hasn’t been used as often.

It is possible to extract important data from News Articles and use that data to gain insight into a specific topic on interest.